**IQR**

**Reference:**

1. **https://medium.com/@pp1222001/outlier-detection-and-removal-using-the-iqr-method-6fab2954315d**

**Significance of outliers:**

* Outliers badly affect the mean and standard deviation of the dataset. These may statistically give erroneous results.
* Most machine learning algorithms do not work well in the presence of outliers. So it is desirable to detect and remove outliers.
* Outliers are highly useful in anomaly detection like fraud detection where the fraud transactions are very different from normal transactions.

**What is IQR Method?**

The IQR method of [outlier detection](https://builtin.com/data-science/how-find-outliers-examples) is a method that dictates that any data point in a [boxplot](https://builtin.com/data-science/boxplot) that’s more than 1.5 IQR points below the first [quartile data](https://builtin.com/data-science/how-to-calculate-quartiles) or more than 1.5 IQR points above the third quartile data is considered an outlier.

**Why Do You Multiply 1.5 in IQR Outlier Detection? Explained**

The interquartile (IQR) method of outlier detection uses 1.5 as its scale to detect outliers because it most closely follows Gaussian distribution. As a result, the method dictates that any data point that’s 1.5 points below the lower bound quartile or above the upper bound quartile is an outlier.

**How to understand a boxplot/ whiskers plot?**

A boxplot tells us, more or less, about the distribution of the data. It gives a sense of how much the data is actually spread out, what its range is and its skewness. It tells us about the various metrics of  data arranged in ascending order.

A screenshot of a computer screen

Description automatically generated

This is the bloxplot. You can plot a boxplot for any numerical column.boxplot contain percentiles of 25, 50 (median), 75, and 100.

* 25th percentile (also known as the first quartile): This means that 25% of the data values are less than or equal to a particular value.
* 50th percentile (also known as the median): This means that 50% of the data values are less than or equal to a particular value. It’s the middle value when the data is sorted in ascending order.
* 75th percentile (also known as the third quartile): This means that 75% of the data values are less than or equal to a particular value.
* 100th percentile: This means that there is no value in the dataset that is greater than or equal to a particular value. In other words, the value at the 100th percentile is the maximum value in the dataset, and no data point in the dataset exceeds or equals this value.

**Steps for IQR method:**

1. The data should be sorted in ascending order
2. Find the IQR using Q1 and Q3

Q1: 25percentile of data

Q2: 50percentile of data

Q3: 75percentile of data

If a dataset has *2n or 2n+1* data points, then  
Q2 = median of the dataset.  
Q1 = median of n smallest data points.  
Q3 = median of n highest data points.

*IQR = Q3 – Q1*

1. To identify the outliers we define the new range by creating 2 boundaries:

Lower Bound: Q1–1.5 \* IQR

Upper Bound: Q3 + 1.5 \* IQR

These boundaries help us determine which data points might be outliers.

1. Deciding whether to remove, transform or replace the outliers

**Any data point that falls below the lower bound (Q1–1.5 \* IQR) is considered an outlier.** These values are significantly lower than the majority of the dataset and are potential candidates for removal or further investigation.

**Conversely, any data point that exceeds the upper bound (Q3 + 1.5 \* IQR) is also considered an outlier**. These values are much higher than the majority of the dataset and may warrant special attention.

Advantages of IQR:

One advantage of the IQR method is that it is robust to skewed data distributions. It identifies outliers based on percentiles, making it less sensitive to extreme values.

Note: Skewed data is data that creates an uneven curve distribution on a graph. We know data is skewed when the statistical distribution’s curve appears distorted to the left or right.

Isolation Forest

Reference:

<https://medium.com/@corymaklin/isolation-forest-799fceacdda4>

Isolation Forest is an unsupervised machine learning algorithm for anomaly detection. As the name implies, Isolation Forest is an ensemble method (similar to random forest). In other words, it use the average of the predictions by several decision trees when assigning the final anomaly score to a given data point. Unlike other anomaly detection algorithms, which first define what’s “normal” and then report anything else as anomalous, Isolation Forest attempts to isolate anomalous data points from the get go.